Supplementary Technical Report for "Analyzing Length-biased Data with Semiparametric Transformation and Accelerated Failure Time Models"

1 Asymptotic Properties of $\hat{\boldsymbol{\alpha}}$

Let $\boldsymbol{\alpha}_0$ be the true value of the regression coefficient vector under the AFT model. We impose the following regularity conditions for a rigorous justification of the asymptotic properties of $\hat{\boldsymbol{\alpha}}$:

(a) Z is a $p \times 1$ vector of bounded covariates, not contained in a $(p-1)$-dimensional hyperplane;

(b) $\sup[t : Pr(V > t) > 0] \geq \sup[t : Pr(C > t) > 0] = t_0$, and $Pr(\delta = 1) > 0$;

(c) $\Gamma_A \equiv -\lim_{n\to\infty} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{q(z_i)\delta_i z_i^{\otimes 2}}{\hat{w}(Y_i)} \right\}$ is nonsingular;

(d) $\int_0^{t_0} [\{\int_t^{t_0} S_C(u)du\}^2 / \{S_C^2(t) S_V(t)\}] dS_C(t) < \infty$;

(e) $E\left[\left\{\delta Z(\log Y - Z^T\boldsymbol{\alpha}_0)\right\} / \{w(Y)\}\right]^2 < \infty$;

(f) $\int_0^{t_0} D^2(s) / \{S_C^2(s) S_V(s)\} \, dS_C(s) < \infty$,
where $D(t) = E\left[q(Z)\left\{\delta Z I(Y \geq s) \int_t^Y S_C(u)du(\log Y - Z^T\boldsymbol{\alpha}_0)\right\} / \{w^2(Y)\}\right]$.

We can establish the consistency of $\hat{\boldsymbol{\alpha}}$ under regularity conditions (a)-(c) as follows. First, we can show that $U_A(\boldsymbol{\alpha})$ has a unique solution $\hat{\boldsymbol{\alpha}}$ since

$$\Gamma_n(\boldsymbol{\alpha}) = dU_A(\boldsymbol{\alpha})/d\boldsymbol{\alpha} = -\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{q(z_i)\delta_i z_i^{\otimes 2}}{\hat{w}(Y_i)} \right\}$$

is negative semi-definite. With probability one, the quantity $n^{-1}U_A^T(\boldsymbol{\alpha})(\boldsymbol{\alpha_0} - \boldsymbol{\alpha})$ converges to

$$\int_z \frac{q(z) z^T z (\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T (\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})}{\mu(z)} dF(z).$$

Then the consistency of $\hat{\boldsymbol{\alpha}}$ follows from the fact that the above limit is non-negative and is zero if and only if $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$.

The derivation of the weak convergence $\sqrt{n}(\hat{\boldsymbol{\alpha}} - \alpha_0)$ can be obtained by the Taylor series expansion of $U_A(\hat{\boldsymbol{\alpha}})$ and the weak convergence of $n^{-1/2}U_A(\boldsymbol{\alpha}_0)$. By Taylor series expansion,

$$\frac{1}{\sqrt{n}}U_A(\hat{\boldsymbol{\alpha}}) = \frac{1}{\sqrt{n}}U_A(\boldsymbol{\alpha}_0) - \frac{1}{n}\Gamma_n(\boldsymbol{\alpha}_0)\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + o_p(1),$$

where $\Gamma_n(\boldsymbol{\alpha}_0)$ is the first derivative of $U_A(\boldsymbol{\alpha}_0)$ and $\frac{1}{n}\Gamma_n(\boldsymbol{\alpha}_0)$ converges in probability to the Hessian matrix of the $U_A(\boldsymbol{\alpha}_0)$, Γ_A. Using the uniform consistency of $\hat{w}(t)$ to $w(t)$, we have

$$n^{-1/2}U_A(\boldsymbol{\alpha}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} q(z_i)\delta_i z_i \frac{\left(\log Y_i - z_i^T\boldsymbol{\alpha}_0\right)}{w(Y_i)}\left\{1 + \frac{w(Y_i) - \hat{w}(Y_i)}{w(Y_i)}\right\} + o_p(1). \tag{1.1}$$

Following from a martingale integral representation for $\sqrt{n}(\hat{w}(t) - w(t))$ by Pepe and Fleming (1989, 1991), we can re-express $\sqrt{n}(\hat{w}(t) - w(t))$ as a martingale integral via integration by parts

$$\begin{aligned} \sqrt{n}(w(Y_i) - \hat{w}(Y_i)) &= n^{-1/2}\sum_{k=1}^{n}\int_0^{Y_i}\left[\int_t^{Y_i} S_C(u)du\right]\frac{dM_k(t)}{\pi(t)} + o_p(1) \\ \sqrt{n}\frac{w(Y_i) - \hat{w}(Y_i)}{w(Y_i)} &= n^{-1/2}\sum_{k=1}^{n}\int_0^{\infty}\frac{h_i(t)}{\pi(t)}dM_k(t) + o_p(1) \end{aligned} \tag{1.2}$$

where $h_i(t) = I(t \leq Y_i)\left[\int_t^{Y_i} S_C(u)du\right]/w(Y_i)$, $\pi(t) = S_C(t)S_V(t)$, $M_k(t) = I(Y_k - A_k \leq t, \Delta_k = 0) - \int_0^t I(Y_k - A_k \geq u)d\Lambda_c(u)$ is the martingale for the residual censoring variable, and $\Lambda_c(u)$ is the corresponding cumulative hazard function. The above martingale integral representation (1.2) implies that

$$\begin{aligned}
& \frac{1}{\sqrt{n}}\sum_{i=1}^{n} q(z_i)\delta_i z_i \frac{\left(\log Y_i - z_i^T \boldsymbol{\alpha}_0\right)}{w(Y_i)} \frac{w(Y_i)-\hat{w}(Y_i)}{w(Y_i)} \\
= \; & \frac{1}{n}\sum_{i=1}^{n} q(z_i)\delta_i z_i \frac{\left(\log Y_i - z_i^T \boldsymbol{\alpha}_0\right)}{w(Y_i)} \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\int_0^\infty \frac{h_i(t)dM_j(t)}{\pi(t)} + o_p(1)
\end{aligned}$$

Note that as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} q(z_i)h_i(t)\delta_i z_i \frac{\left(\log Y_i - z_i^T \boldsymbol{\alpha}_0\right)}{w(Y_i)} \to D(t).$$

Therefore,

$$n^{-1/2}U_A(\boldsymbol{\alpha}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{ q(z_i)\delta_i z_i \frac{\left(\log Y_i - z_i^T \boldsymbol{\alpha}_0\right)}{w(Y_i)} + \int_0^\infty \frac{D(t)dM_i(t)}{\pi(t)} \right\} + o_p(1).$$

Hence, under regularity conditions (d)-(f), $n^{-1/2}U_A(\boldsymbol{\alpha}_0)$ is asymptotically normally distributed by the Central Limit Theorem. This, combined with an application of Slutsky's theorem, implies that $\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)$ converges weakly to a normal distribution with mean zero and variance-covariance matrix $\Gamma_A^{-1}\Sigma_A\Gamma_A^{-1}$, in which Σ_A is the asymptotic variance-covariance matrix of $n^{-1/2}U_A(\boldsymbol{\alpha}_0)$.

2 Asymptotic Efficiency of Two Approaches under AFT Model

Based on the joint distribution of (A, Y) and C conditional on covariates Z,

$$\begin{aligned}
& E\left\{ \frac{\delta(\log Y - Z^T\boldsymbol{\alpha})}{Y S_C(Y-A)} | Z = z \right\} \\
= \; & E\left\{ \frac{1}{\mu(z)} \int_0^\infty \int_0^y f_U(y|Z=z)S_C(y-a)\frac{(\log y - z^T\boldsymbol{\alpha})}{yS_C(y-a)} da dy \right\} \\
= \; & E\left\{ \frac{1}{\mu(z)} \int_0^\infty f_U(y|Z=z)(\log y - z^T\boldsymbol{\alpha})dy \right\} = 0.
\end{aligned}$$

Accordingly, an alternative asymptotic unbiased estimating equation for $\boldsymbol{\alpha}$ can be constructed as

$$U_S(\boldsymbol{\alpha}) = \sum_{i=1}^{n} q(z_i)\delta_i z_i \frac{\left(\log Y_i - z_i^T\boldsymbol{\alpha}\right)}{Y_i\hat{S}_C(Y_i - A_i)} = 0, \tag{2.3}$$

where q is a positive, scalar weight function. The estimating equation leads to a closed-form solution for $\boldsymbol{\alpha}$,

$$\hat{\boldsymbol{\alpha}}_S = \left\{\sum_{i=1}^{n} \frac{q(z_i)\delta_i z_i z_i^T}{Y_i\hat{S}_C(Y_i - A_i)}\right\}^{-1} \sum_{i=1}^{n} \frac{q(z_i)\delta_i z_i \log Y_i}{Y_i\hat{S}_C(Y_i - A_i)}.$$

Let $\boldsymbol{\alpha}_0$ be the true value of the regression coefficient vector. We can prove that the estimating equation $U_S(\boldsymbol{\alpha})$ yields a unique and consistent estimator $\hat{\boldsymbol{\alpha}}_S$ under some regularity conditions. Moreover, $\sqrt{n}(\hat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_0)$ converges weakly to a normal distribution with mean zero and variance-covariance matrix $\Gamma_S^{-1}\Sigma_S\Gamma_S^{-1}$, in which Γ_S is the Hessian matrix of the $U_S(\boldsymbol{\alpha}_0)$ and Σ_S is the asymptotic variance-covariance matrix of $n^{-1/2}U_S(\boldsymbol{\alpha}_0)$.

In contrast, our proposed estimating equations $U_A(\boldsymbol{\alpha})$ use an inverse of the integral of the Kaplan-Meier estimator as the weight,

$$U_A(\boldsymbol{\alpha}) = \sum_{i=1}^{n} q(z_i)\delta_i z_i \frac{(\log Y_i - z_i^T\boldsymbol{\alpha})}{\hat{w}(Y_i)} = 0. \tag{2.4}$$

While the two estimating equations are both valid for large sample properties, an interesting question is which estimating equation leads to a more efficient estimator of $\boldsymbol{\alpha}$, and under what conditions. In this section, we study the difference between the two asymptotic variance-covariance matrices if the censoring distribution is known,

$$Var(\hat{\boldsymbol{\alpha}}_S) - Var(\hat{\boldsymbol{\alpha}}) = \Gamma_S^{-1}\Sigma_S\Gamma_S^{-1} - \Gamma_A^{-1}\Sigma_A\Gamma_A^{-1},$$

where Σ_S and Σ_A denote the variance-covariance matrices of $n^{-1/2}U_S(\boldsymbol{\alpha}_0)$ and $n^{-1/2}U_A(\boldsymbol{\alpha}_0)$ respectively. Note that for any censoring distribution, the two Hessian matrices Γ_S and Γ_A are the same, since

$$\Gamma_S \;=\; E\left[E\left\{\frac{q(Z)\delta ZZ^T}{YS_C(Y-A)}\Big| Z\right\}\right] = E\left\{\frac{q(Z)ZZ^T}{\mu(Z)}\right\}$$

and

$$\Gamma_A = E\left[E\left\{\frac{q(Z)\delta ZZ^T}{w(Y)}\Big|Z\right\}\right] = E\left\{\frac{q(Z)ZZ^T}{\mu(Z)}\right\}.$$

It is then essential to compare the difference between the variance-covariance matrices Σ_S and Σ_A. We first show that the covariance matrix of $n^{-1/2}U_S(\boldsymbol{\alpha}_0)$ and $n^{-1/2}U_A(\boldsymbol{\alpha}_0)$ is equal to the variance-covariance matrix Σ_A,

$$\begin{aligned}
& Cov\left(n^{-\frac{1}{2}}U_S(\boldsymbol{\alpha}_0), n^{-\frac{1}{2}}U_A(\boldsymbol{\alpha}_0)\right) \\
= \; & E\left[q(Z)^2 ZZ^T E\left\{\frac{\delta(\log Y - Z^T\boldsymbol{\alpha}_0)^2}{YS_C(Y-A)\int_0^Y S_C(t)dt}\Big|Z\right\}\right] \\
= \; & E\left[\frac{q(Z)^2 ZZ^T}{\mu(Z)}\int\int_0^y \frac{(\log y - Z^T\boldsymbol{\alpha}_0)^2}{yS_C(y-a)\int_0^y S_C(t)dt} S_C(y-a)f_U(y|Z)dady\right] \\
= \; & E\left\{\frac{q(Z)^2 ZZ^T}{\mu(Z)}\int\frac{(\log y - Z^T\boldsymbol{\alpha}_0)^2}{\int_0^y S_C(t)dt} f_U(y|Z)dy\right\} = \Sigma_A. \qquad (2.5)
\end{aligned}$$

Because the variance-covariance matrix $Var\left(n^{-\frac{1}{2}}U_S(\boldsymbol{\alpha}_0) - n^{-\frac{1}{2}}U_A(\boldsymbol{\alpha}_0)\right)$ is non-negative definite, with equation (2.5) we can ensure that the following difference in variance-covariance matrixes is always non-negative definite,

$$\Sigma_S - \Sigma_A = \Sigma_S + \Sigma_A - 2Cov\left(n^{-\frac{1}{2}}U_S(\boldsymbol{\alpha}_0), n^{-\frac{1}{2}}U_A(\boldsymbol{\alpha}_0)\right) = Var\left(n^{-\frac{1}{2}}U_S(\boldsymbol{\alpha}_0) - n^{-\frac{1}{2}}U_A(\boldsymbol{\alpha}_0)\right).$$

Therefore, the estimator obtained from $U_A(\boldsymbol{\alpha})$ is found to be asymptotically more efficient than that from $U_S(\boldsymbol{\alpha})$ under any censoring distribution.